WHITE PAPER

New and improved proteomics technologies for understanding complex biological systems: Addressing a grand challenge in the life sciences

Leroy E. Hood[1], Gilbert S. Omenn[1,2], Robert L. Moritz[1], Ruedi Aebersold[3], Keith R. Yamamoto[4], Michael Amos[5]**, Jennie Hunter-Cevera[6], Laurie Locascio[5] and Workshop Participants****

[1] Institute for Systems Biology, Seattle, WA, USA
[2] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[3] Institute for Molecular Systems Biology, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland
[4] Department of Cellular and Molecular Pharmacology, University of California San Francisco (UCSF), CA, USA
[5] National Institute of Standards and Technology, Department of Commerce, Gaithersburg, MD, USA
[6] Discovery and Analytical Sciences, RTI, International, Research Triangle Park, NC, USA

Received: April 13, 2012
Revised: June 6, 2012
Accepted: June 20, 2012

This White Paper sets out a Life Sciences Grand Challenge for Proteomics Technologies to enhance our understanding of complex biological systems, link genomes with phenotypes, and bring broad benefits to the biosciences and the US economy. The paper is based on a workshop hosted by the National Institute of Standards and Technology (NIST) in Gaithersburg, MD, 14–15 February 2011, with participants from many federal R&D agencies and research communities, under the aegis of the US National Science and Technology Council (NSTC). Opportunities are identified for a coordinated R&D effort to achieve major technology-based goals and address societal challenges in health, agriculture, nutrition, energy, environment, national security, and economic development.

Keywords:
Complex systems / Democratization of proteomics / Economic growth / Grand challenges / Integration / Systems biology

1 Introduction

1.1 Aim of this White Paper

The aim of this White Paper, based on the 2011 Gaithersburg Workshop, is to identify new proteomics analytical tools and applications that can bring broad benefits to the biosciences and to the US economy. This report highlights opportunities for multiple US Government agencies to pursue their priorities with a coordinated R&D effort that integrates proteomics into major life sciences initiatives. The prospects for budget synergies for the agencies and for a large multiplier in economic activity and job growth in the relevant applied biotechnology sectors can make this opportunity attractive even in difficult budget times. A 2011 Battelle report [1] estimated an $800 billion economic impact over 22 years from the Nation's investment in genomics and biotechnology; proteomics can play an analogous role in catalyzing economic benefits for the next stage of biotechnology, particularly since proteins are two steps closer than genes to most biological phenomena and diseases. In addition, this initiative would enhance

Correspondence: Dr. Gilbert S. Omenn, University of Michigan, 2065 Palmer Commons, Ann Arbor, MI 48109–2218, USA
E-mail: gomenn@umich.edu
Fax: +1 734-615-6553

Abbreviations: NCI, National Cancer Institute; **NIST**, National Institute of Standards and Technology; **NSTC**, National Science and Technology Council; **SNP**, single-nucleotide polymorphisms; **SRM**, selected reaction monitoring

*Additional corresponding author: Dr. Leroy E. Hood, E-mail: lhood@systemsbiology.org
**The Workshop and this White Paper are dedicated to the memory of Dr. Michael Amos of NIST. His vision to organize this effort and his encouragement to always push the boundaries of science will long be remembered.
***See the Addendum, for the full list of additional Workshop participants.

the role of US scientists, institutions, and companies in a growing, high-profile international endeavor.

1.2 Vision of a grand challenge for proteomics

The vision of a grand challenge for proteomics is to make bold advances in utilizing and advancing the technology platforms and knowledge bases for quantifying and characterizing proteins in functional protein networks, thereby facilitating societally important applications in health, agriculture, nutrition, energy, environment, and national security.

The proteome is the operating system for nearly all biological functions. It is the link between the genome and phenotypes. It undergoes dynamic changes in different cells and organs, during development, in response to environmental stimuli, and in disease processes. Understanding the dynamics of protein interactions with other proteins, nucleic acids, and metabolites is the key to delineating biological mechanisms and understanding disease.

1.3 Background for the workshop

In 2009, the Office of Science and Technology Policy issued a call for grand challenges in biotechnology. National Institute of Standards and Technology (NIST) submitted a proposal for a workshop and white paper on new technologies for proteomics, which was approved by the National Science and Technology Council (NSTC) Biotechnology Subcommittee. The Workshop convened by Michael Amos of NIST, Leroy Hood of the Institute for Systems Biology, Ruedi Aebersold of ETH-Zurich, and Keith Yamamoto of UCSF was held on 14–15 February 2011 in Gaithersburg, MD. Attendees are listed in the Appendix (additional participants) and as coauthors above.

Further stimulus for this Workshop came from recent reports from the National Research Council on "A New Biology for the 21st Century" [2] and "Research at the Intersection of Physical and Life Sciences" [3]. Those reports identified a need for technologies to understand biological systems in sufficient depth to fulfill societal goals of advancing and protecting health, bioenergy, the environment, food production, green manufacturing, and national security. These reports identified the need for new research approaches for studying biological systems that bring together the physical, chemical, biological, and computational sciences.

1.4 Baseline for the grand challenge for proteomics

There have been dramatic advances in the past 5 years in identification and quantification of proteins in biological systems, generating confidence that a complete parts list of the primary products of the 20 300 protein-coding genes in humans and corresponding whole proteomes of other organisms is within reach [4]. The targeted MS approach of Selected Reaction Monitoring (SRM) led by American and Swiss scientists has produced the SRM Atlas with spectra for approximately six informative peptides (signatures) for each of the 20 300 human protein gene products and a corresponding atlas for all of the 6500 yeast protein products [5, 6]. For accurate quantification, stable heavy isotope-labeled peptide reagents are now readily available for all of these peptides through commercial sources. For increased signature peptide detection, antipeptide capture reagents can reduce the concentrations for the limits of detection by current instrumentation by two orders of magnitude [7].

Remarkable advances in MS instruments now permit identification of 7000 to 10 000 proteins in single studies of cultured human cell lines and other specimens with the latest Orbitrap hybrid mass spectrometers [8, 9]. The detection of dozens to hundreds of peptides simultaneously by SRM coordinates, derived from the SRM Atlas and applied to nonscanning triple-quadruple MS instruments from multiple manufacturers, provides unprecedented targeted quantitative analysis. On the horizon is the enhancement of top-down MS that yields sequences of proteins (currently up to about 50 kDa molecular mass) without requiring the variable steps of proteolytic digestion and peptide fractionation. US manufacturers are at the forefront of many of these instrument advances.

Due to PTMs, single-nucleotide polymorphisms (SNPs), and alternative splicing of proteins, there are numerous different structures of individual primary products, making the dimensions of this grand challenge much larger than the human genome sequencing project. PTMs can be detected with Electron Transfer Dissociation (ETD) collision spectrometry [10]. Genomics alone cannot interrogate and delineate this protein diversity and its functional consequences.

The Human Protein Atlas, led by Uhlen et al. in Sweden, is a great example of integrating genomics and proteomics, using predicted protein sequences and epitopes to produce antibodies that can capture proteins and map by immunohistochemistry the expression and intracellular localization of each specific protein in 46 tissues of humans. The Protein Atlas now covers 12 238 proteins and is projected to reach 14 000 of the 20 300 proteins with version 10 in September 2012 [11]. The antibodies developed in this effort are reagents applicable to many biological studies. For example, these protein epitope signature tagged peptides (PrESTs) have been labeled with stable isotopes and spiked into cell lysates to facilitate absolute quantitation of the targeted proteins with MS [12].

Finally, a well-coordinated, global, curated data and knowledge base for proteomics findings is essential for broad dissemination and secondary analyses. Such a knowledge base, covering about 13 000 proteins, has been created through sustained major European funding for the Swiss Institute for Bioinformatics (SwissProt, NeXtProt) and the European Bioinformatics Institute (UniProt, PRIDE) and through US efforts that provide standardized reanalysis of MS-based proteomes by the Institute for Systems Biology

in Seattle (Peptide Atlas, SRMAtlas, PASSEL). These global resources are linked through the EU-funded ProteomeXchange Consortium (www.proteomexchange.org). Identifying the remaining proteins, characterizing their many variants due to SNPs, RNA splicing, and PTMs, and integrating proteomic and genomic analyses through biological networks and protein and RNA complexes is truly a feasible grand challenge.

Meanwhile, the international Human Proteome Organization (HUPO) has stimulated and led ten initiatives on organ proteomes (liver, brain, kidney, and heart), biofluid proteomes (plasma/ serum [13], urine, and cerebrospinal fluid), model organism proteomes, protein standards, and antibodies. In 2010, HUPO announced a global Human Proteome Project in which numerous countries have stepped forward to lead efforts focused on the proteins coded for by genes on individual chromosomes and on protein interactions and networks that mediate a wide range of biological and disease processes [14, 15]. An NSTC initiative on bold new technologies represents a distinctive opportunity for global leadership by the United States that would be a major contribution to the Human Proteome Project.

Apart from multiyear NIH center grant programs renewed in 2009 by the National Heart, Lung, and Blood Institute (NHLBI) and in 2011 by the National Cancer Institute (NCI) for Clinical Proteomics Technologies Consortia, proteomics support in the United States is mostly embedded in applied biology project grants under the NCI Early Detection Research Network and clinical proteomics programs of the NHLBI, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), and other institutes and agencies. The Workshop participants concluded that the time is ripe to develop and apply bold advances in next-generation proteomics technology platforms with specificity, depth, and quantification to characterize the nature and function of the human proteome. Also in late 2011, the NIH convened a Human Proteome Workshop that focused on protein interactomes, biological networks, and the path to clinically useful biomarkers [16].

2 Linking the genome to normal and disease phenotypes: The key role of proteins in realizing the full potential of the human genome project

The Human Genome Project has transformed many aspects of human biology and medical research. It was made feasible by audacious goals and successful new technology platforms. It overcame severe initial skepticism about feasibility and payoff. The essential first stage of the Human Genome Project was to design and use potent new technology platforms for sequencing and synthesizing DNA and protein molecules [17]. This radically changed the study of genes and gene regulation in all organisms. The Genome Project democratized genetics, making all genes accessible to biologists everywhere; generated a comprehensive "parts list" of genes and, by inference, their protein products; stimulated a holistic approach to studying biological complexity through systems biology; catalyzed the emergence of a whole new commercial sector with high-throughput instruments, reagents, services, and products; pioneered the large-scale application of computer science and informatics to biology; demonstrated the power of open-source data and data resources; produced the first rigorous standards for biological data; facilitated access to the genomes of plants, animals, and microbes with stunning new findings; revolutionized our thinking of evolution through comparative genomic studies; and transformed our thinking about medical diagnostics and targeted therapies. Nevertheless, there are many scientists, public figures, and journalists concerned about the slow progress in transforming medicine and public health.

By analogy, the grand challenge in Proteomics can have spectacular results similar to those of the genome project – democratizing the study of proteins, creating a broader and deeper parts list for systems biology, characterizing protein interactions and biological networks that mediate physiological and pathological processes, building an economic engine for instruments and reagents and omics-based tests, and scientifically linking all species to address societal goals for health, food, energy, environmental sustainability, and national security. Based on successes in many fields, the enunciation of a "grand challenge" can actually contribute to its solution, by stimulating essential technology development and by identifying presently inaccessible gaps in technologies [18].

An important point for emphasis is a holistic or systems approach to deciphering the complexity of biology and disease. A simple analogy for systems thinking is that of understanding how a radio converts electromagnetic waves into sound waves. This could proceed in three steps: first, taking the radio apart and cataloguing all of the components; second, determining the functions of the individual parts; and, third, putting the components together in circuits (networks) and learning how the networks collectively contribute to converting radio waves into sound waves. The Genome Project identified all genes and by inference all proteins. For the last 40 years, biologists have studied the individual components of life (genes and proteins). The goal now is to understand how these proteins and genes are integrated into the biological networks and molecular machines that function to convert the information of the genome and the environment into the phenotype of the organism. A critical aspect of systems approaches is deciphering biological and medical complexity by following the dynamics of these interactions, in part with technologies that can identify and quantify all proteins, protein isoforms, and protein interactions of living organisms.

We are at an inflection point for proteomics technologies. As with the early years of the Human Genome Project, we can "catch the wave" and create a transformational future. This mission goes far beyond creating a parts list of all proteins.

2.1 Proteins

Proteomics technology development and applications can dramatically augment genomic efforts. The proteome arises as a result of an enormous amplification of complexity that occurs stepwise in the biological translation of information from DNA to RNA to proteins. Proteins are:

(i) the effector molecules that execute many critical functions of life, whose lesions lead to diseases.
(ii) the major components of biological networks that control many functions of cells and their intercellular interactions, by capturing, transmitting, and integrating biological information and passing it on to molecular machines.
(iii) the cell-surface receptors and channels that sense the environment and bring this information into living cells, with collisions between the digital information of the genome and the diverse insults from environmental exposures determining susceptibility or resistance to various diseases.
(iv) the regulators and downstream mediators of gene expression and biological networks.
(v) the targets of nearly all drugs and, increasingly, themselves effective drugs.

2.2 Complexity of the proteome

The 20 300 protein-coding genes in humans give rise to perhaps a million different protein isoforms in the human proteome. While there are two nearly identical copies of every autosomal gene in all nucleated cells, there may be anywhere from zero to millions of copies of a protein in a given cell, representing an enormous range of concentrations, which can vary strikingly in response to internal and external stimuli. Thus, quantitative real-time measurements are essential for assessing the dynamics of the proteome. Each gene-encoded protein has numerous isoforms that reflect mutations, gene fusions, alternative splicing, mRNA editing, chemical modifications, and proteolytic processing of the proteins. Ultimately, we must develop technology platforms to identify, quantify, and determine the functions of these many protein isoforms.

Proteins interact with other proteins, nucleic acids, lipids, and other small molecules, including tightly bound metal atoms, to form molecular machines mediating movement or ion flow, membrane structures, large interactive biological networks, and gene regulatory complexes. Finally, proteins are dynamic in ways genes are not: they can undergo rapid changes in 3D shape, change locations within a cell, change rates of secretion or release to the circulating blood and lymphatics, and function in molecular machines and biological networks. Thus, proteins represent multidimensional, dynamic, analog information rather than linear, mostly static digital information.

The Grand Challenge for Proteomics is to develop technologies and research resources capable of identifying and characterizing these molecular species and following their dynamic molecular interactions in health and disease.

3 Emerging proteomics technologies that jump-start the grand challenge

3.1 MS

MS methods currently are the backbone of experimental proteomics for global protein analyses. An impressive, rapidly expanding array of instruments matched to particular complementary applications has galvanized progress, as noted previously [4]. MS-based proteomics was greatly accelerated by information from the Human Genome Project for sequence matching. The goals for innovation are: to markedly increase signal/noise in identifying and sequencing peptides, to detect and quantify specific peptides with PTMs, SNPs, or splicing, and to greatly increase the throughput to make assays useful for clinical studies and population studies. Another critical area is to be able to follow closely the dynamics of how proteomes change (in concentration and in structure) in response to environmental signals and disease. Protein-capture reagents can recognize and pull down targeted proteins and multimolecular complexes, followed by identification with MS.

Proteins emerging from biomarker discovery experiments can be assayed through distinctive peptides with SRM. A key element is greater sensitivity of identification of proteins in complex mixtures like tissue lysates and plasma. An exciting example is SWATH, Sequential Window Acquisition of all Theoretic Mass Spectra. This approach records a permanent digital record of fragment ion spectra of each object in a sample, which can be identified from the corresponding reference spectrum in the SRM Atlas. SWATH maintains the dynamic range of SRM but vastly increases its coverage per unit time, essentially becoming a protein microarray [19]. This technique will be powerful in more comprehensively assessing proteome dynamics – a critical need of contemporary proteomics and systems biology.

3.2 Protein-capture reagents

Protein-capture reagents can identify, quantify, localize, and purify specific proteins for a wide variety of studies. Promising reagents include polyclonal and monoclonal antibodies, aptamers, peptides, and DNA–peptide hybrids. The desired attributes are highly avid, highly specific even in complex specimens, renewable, scalable, robust, stable in storage, easy to mass-produce with high quality, easy to transfer, relatively inexpensive, and with well-defined uses. A special attribute of certain emerging reagents is bifunctional chemistry that

enables target molecules to be detected in vivo through the protein-capture agents and associated reporter groups.

An advance based on click chemistry is the generation of multivalent peptide reagents [20]. A large D-amino acid peptide library of 6-mers is screened against the protein to be captured to identify low-affinity anchor peptides. Additional low-affinity peptides are identified similarly. The anchor peptide and additional peptides are linked in three-dimensional orientations by click chemistry to form a dimer and then a trimer, which can have affinities and specificities that match or exceed the best monoclonal antibodies. These agents are also useful for in vivo imaging. Recently, these reagents have been directed at specific epitopes on proteins, such as a phosphorylated versus a nonphosphorylated site. Hence, they can follow important biological modifications of proteins and have the potential to become effective drugs.

Surface-based bioaffinity measurements utilizing nanoparticles, aptamers, enzymatic transformations, and intracellular signal amplification methods can impart higher sensitivity and selectivity to analyses of complex protein networks.

3.3 In vivo molecular imaging

Protein-capture agents can be designed for delivery to target tissues in model organisms or humans for physiological, diagnostic, and therapeutic purposes. Imaging is a key application for protein-capture agents. Visualization can be mediated by fluorescence, luminescence, radioactivity, or other labeling methods, and can be combined with nanotechnology for carriers. The key to understanding many physiological and disease processes is to follow the dynamics of molecular transitions in vivo. Small, high-affinity/high-specificity reagents with appropriate reporter groups are essential for such imaging technologies. Use of fluorescence recovery after photobleaching (FRAP) has provided measurements of proteins and their complexes down to single-molecule sensitivity [21]. Greater diversity of imaging technologies will be helpful. Next-generation imaging tools will characterize complex intracellular structures, such as virus capsid shells and ion channel membrane pores. They will help visualize complex protein interactions and possibly distinguish differentially expressed isoforms of target proteins.

3.4 Ultrasensitive single-cell and single-molecule analyses of proteins

Single-cell analyses are becoming a fundamental approach in cell biology. The key technology platforms utilize powerful adaptations of microfluidics, nanotechnology, and new chemistries. For example, Rissin et al. developed a single-protein molecular detection strategy by "singulating" enzyme-linked protein molecules on microspheres in arrays of 50-femtoliter reaction chambers with digital readout of fluorescence, and demonstrated this highly sensitive approach on prostate-specific antigen and tumor necrosis factor-alpha in serum [22]. Single-cell and ultrasensitive protein analyses will enable investigations of the differential roles of distinct cells in complex mixtures in tissues or blood.

3.5 In silico protein folding approaches

A rapidly growing list of diseases caused by misfolding of proteins (from prion brain disease and Alzheimer disease to diabetes [23]) puts a focus on connecting proteomics with structural biology. Three-dimensional structures can provide fundamental insights into the protein functions not only of humans, but also plants, animals, and microbes, bridging protein chemistry and proteomics. For instance, computational modeling of conformational changes due to exon swapping in pairs of differentially expressed protein splice variants in cancers has shown the power of modeling and inference for functional consequences [24]. In silico protein folding may also be important for determining whether gene variants identified in the analyses of human genome sequences lead to proteins whose structures are sufficiently altered to become nonfunctional.

3.6 Computational visualization and integration of molecular findings across the omics platforms

The goal is to generate predictive, actionable, testable models of biological and disease processes and responses to external stimuli. This requires connecting and integrating genomic, epigenomic, transcriptomic, proteomic, metabolomic, and many types of phenotypic data. In taking a systems approach to understanding biological processes, it is essential first to identify the relevant parts list; then one must integrate these different types of "parts" into descriptive, graphical, or mathematical models for biological networks that give insights into how their dynamic behavior captures, transmits, integrates, and passes on information to the molecular machines that execute the functions of life. Tools such as Cytoscape and its many plug-ins facilitate graphical presentation of such complex information and molecular relationships.

3.7 Computational and mathematical methods and models

Validation and quantification of the data and insights from proteomics and other omics technologies depend on computational sciences and scientists capable of bridging the biosciences and bioinformatics. At the highest level, there is the question of how one turns data into knowledge. One of the grand challenges in biological measurements is to validate the quality of the data and to be able to deal with the tremendous signal-to-noise and overfitting challenges that come with

large data sets [25]. We also need to think about how to effectively capture, store, mine, integrate, and finally model these data – so that predictions and actionable consequences result. We need to create the tools that will let biologists and medical researchers utilize the power of proteomics through assays that cover relevant biological networks and make it easy to follow their dynamics. Novel algorithms for statistical models are needed to decrease false-positive peptide and protein identifications, increase signal/noise ratios, sequence longer peptides, and characterize high-charge states.

Computation is transforming biology, just as it has transformed many other fields. Data dimensionality is enormous and growing rapidly; in 10 years, we may have all kinds of biological systems and even individual patients with billions of data points to be linked through genotype–phenotype correlations.

4 Measurable goals and 5-year deliverables from a grand challenge for proteomics technologies

In setting goals and deliverables, let us recapitulate the key features of proteins for global analysis of the proteome: (1) while genes are digital in nature with a four-letter language, proteins are analog with a 20 letter language – genes operate in a one-dimensional world and proteins in a three-dimensional world; (2) proteins lack the molecular complementarity of DNA and hence cannot be amplified prior to measurement – thus, we must develop ultrasensitive techniques to measure and analyze a few or single protein molecules; (3) proteins have extreme complexity due to modifications by gene mutation, RNA editing, RNA splicing, up to 400 types of covalent changes, and protein processing; (4) proteins are dynamical, changing their three-dimensional structures, positions in the cell, concentrations at different cellular sites, sequences, covalent chemistries, and interactions with other proteins and molecules of many types in response to endogenous and exogenous stimuli; (5) proteins exhibit a 10^6 dynamic range in tissues and a 10^{12} dynamic range in blood, making quantification essential; (6) there are two fundamental ways of identifying and quantifying proteins: (a) MS and (b) protein-capture agents (e.g. antibodies); and (7) we ultimately need to be able to measure proteins in all of their dimensions in the context of the single cell, the fundamental unit of function in living organisms, which will require the development of microfluidic and nanotechnology techniques for handing single cells.

With these specific challenges in mind, here are goals and deliverables we consider feasible over the coming 5-year period:

(i) Enhance the **sensitivity** of MS-based protein identification by 100×–1000× in tissues and plasma to match the most sensitive antibody (ELISA) assays and monitor the dynamics of low-abundance proteins with high biological relevance. This is the scale of technological gain that made the Genome Sequencing Project feasible. The required amount of material for protein analyses will move from 10^6 cells to 10^3 cells, and then toward single cells in concert with additional technologies (below).

(ii) Create MS **reference spectra** for peptides of all proteins, including those with PTMs (glycosylation, phosphorylation, acetylation, ubiquitinylation, many others); gain value from the substantial proportion of high-quality spectra not yet recognized as modified peptides; and expand the existing PeptideAtlas and newly produced SRM Atlas. Make all of these resources publicly available through NIST (as for SpectraST), NIH, and other federal agencies. Connect MS data with protein capture data through the Human Protein Atlas.

(iii) Deploy **microfluidics and nanotechnology** together with proteomics to permit sensitive analyses of single cells and very low-abundance protein molecules and their interactions with other proteins, nucleic acids, and small molecules. These technologies enable miniaturization of sample size, integration of multiple chemical procedures in the assay, parallelization of measurements to increase throughput, and automation of assays.

(iv) Enhance **throughput** of proteomics assays so that reproducible, sensitive results can be obtained on many hundreds of specimens per day, useful for drug development, patient monitoring, epidemiological studies of populations, and complex time-course experiments tied to signaling pathways, biological networks, gene regulation, and disease phenotypes. Ultimately, carry out high throughput proteomics assays by protein-capture agents arrayed on microfluidic chips so that thousands of measurements may be made in a few minutes on samples from a fraction of a droplet of blood (or solubilized tissue). These assays will revolutionize our ability to understand biological functions, present new strategies for diagnostics, open up exciting new possibilities for identifying effective drug targets – and then enable the efficient selection of drugs for these targets.

(v) Move **massive proteomics data resources** to the **cloud** along with the appropriate analytic tools so that very large-scale data analysis may be managed from desktop computers; of course, this requires figuring out how to input these data into the cloud efficiently and rapidly. This capability will link **sustainable data repositories** and proteomics atlases, with open access, and will have data links and analytical tools embedded to facilitate modeling of phenotypic responses.

(vi) Develop **software** that will let any biologist integrate proteomics data into the broad spectrum of omics data (epigenetic, genomic, transcriptomic, miRNA, metabolomic, and interactomic) – to create metadata structures from which predictive models about biological systems can be generated. This begins with

enhanced proteomics workflows to process terabytes of LC/MS-MS data with scaling and high throughput from thousands of samples. A current domain for integrative omics analysis is pathway-based cancer therapies. Another realm for tools that integrate omics data is to characterize organisms from meta-proteomes of complex human microbiomes and the microbial communities that dominate the Earth's biomass.

(vii) Adopt **standardized criteria and establish software for quality assessment** and quality assurance of the various types of proteomics data [26].

(viii) Apply these powerful proteomics approaches to the **proteomes** of **important model organisms** – microorganisms, *Arabidopsis*, *Drosophila*, *Caenorhabditis elegans*, zebrafish, mice, rats, and primates – to enable comparative proteomics.

5 Expected impacts from the grand challenge in proteomics for sectoral S&T goals and for economic growth

5.1 Human health (NIH, lead)

Large-scale analysis of proteins has been inspired by the realization that the proteome is inherently more complex and more relevant to function than the genome alone. Software-driven data analysis and inference is vital for understanding complex diseases, disease prognosis and progression, personalized treatment selection, drug targeting, and drug safety assessments.

Strategies for personalized, predictive, preventive, and participatory (P4) medicine depend on an integrated omics-based R&D approach [27]. Protein and complementary RNA (mRNA, miRNA, lncRNA) molecular signatures for functions and dysfunction of each organ in the body and for major disease processes will contribute importantly to diagnosis, prognosis, and monitoring of therapeutic and preventive interventions. Such biomarkers will help detect early disease, monitor disease progression, stratify patients with a common diagnosis like Alzheimer disease or post-traumatic stress into subtypes for proper impedance match with emerging therapies, and monitor desired and adverse responses to therapy. A pilot example is serum profiling of liver fibrosis due to hepatitis C [28]. The quantification of proteins in complex mixtures will be key to making blood and other biofluids windows for surveying the health and disease status of individual patients. New technologies that can measure more sensitively and more specifically the key isoforms of proteins in tissue specimens or in the circulating plasma will make proteins sensors of our environmental exposures, risk factors, and microbiome influences.

We realize from the extended and continuing period of incubation of genomics into clinical practice that our perspective must extend beyond the first 5 years. Simultaneously, we must learn how to practice P4 Medicine without loss of privacy and confidentiality or excessive cost. In fact, an explicit goal should be to moderate costs through more effective and efficient health care and preventive medicine. Input from the public, federal, and state perspectives will be important to meet these challenges.

5.2 Agriculture and food (USDA, lead)

The Green Revolution of the 20th Century transformed agriculture by scientifically increasing productivity of crops and nutritive value of foods. Those gains hit their peak recently. Fortunately, improvements in remote sensing, transportation, weather prediction, and recombinant DNA technology are transforming agriculture again. This process plays out over decades. Thus, there is a critical need for long-term, sustained investments in agricultural R&D through the US Department of Agriculture, international agencies, and companies. Plant genomics and proteomics are some of the most exciting and productive areas of omics research. The integration of all of the omics including proteomics into predictive models will be key for redesigning plant genes to optimize nutrition, taste, durability, resistance to infectious agents, and other objectives. Detailed knowledge is required for the different plant and animal species and for characterizing adaptations to changing crop conditions, especially in light of climate perturbations. Plants and animals also are excellent models for learning how genetic and environmental inputs are integrated by all living organisms. *Arabidopsis* has become recognized as one of the key model organisms in proteomics and all of biology. Complete SRM assays for all of its proteins will transform our ability to decipher its complexity in the service of improving agricultural practices.

Proteins and protein functions are at the heart of numerous questions about understanding crop and livestock biology, increasing productivity, and enhancing resistance or other adaptations to environmental stressors. Proteins can be biomarkers for many properties of crops and livestock and of the food products throughout the many steps of the food chain to animal and human consumers. Significant human public health goals are to increase the nutritional value of common foods, as well as designing foods to fit specific metabolic disorders (nutrigenomics/nutriproteomics) and to help mitigate the rapidly growing problem of obesity, especially childhood obesity. The understanding of PTMs and protein–protein interactions through network analysis will be crucial to gaining benefits from directed improvements in diets and in food quality.

5.3 Energy and environment (NSF, DoE, EPA, DoD leads)

A very interesting goal for proteomics is to create a "read-out" for the "health" of ecosystems. One baseline is the characterization of microbial communities, which have evolved stunningly diverse capacities for altering the chemical forms of

most elements of the periodic table, selectively interconverting organic molecules, producing valuable products through fermentation, and communicating and competing under diverse conditions. They may be mobilized to break down refractory organic molecules like chlorinated fluorocarbons, lignin, and cellulose; remove actinides like uranium from aquifers or waste streams; and bioleach copper from sulfide minerals. Proteomics capabilities could be tapped to develop alternative bioenergy sources and understand and enhance carbon sequestration. All of these goals require knowledge of the interrelated processes of plants and microbial communities and their links to biogeochemical cycles. Participants in this Workshop were optimistic about applying emerging knowledge of proteins in pathways that might shift the biofuel feedstock to currently intractable parts of cells, thus reducing the impact of biofuel production on available supplies and prices of foods, especially corn. If biofuel production could be scaled sufficiently, with net energy gain, it might help reduce dependence on imported oil.

Another focus is the balance of light utilization and protection against oxidative species in photosynthetic plants and algae; proteomics research is expected to provide insights into photosynthetic processes and carbon, nitrogen, and water cycles. Microbes in the plant environment fix nitrogen, mineralize nutrients from decaying organic matter, scavenge phosphorus, produce plant growth promoters, aid soil structure, and protect against disease agents. Proteomics studies will contribute to characterizing new genomes and their gene products, modeling of stable states, communities, and ecosystems spanning all scales from molecular to global. Proteomics will provide useful knowledge about these and many other features of symbiosis and systems biology.

5.4 National security and counterterrorism (DoD, DHS, NIH)

Advanced proteomics may have special value through the use of proteins in threat detection, prediction, and deterrence. Many national security applications focus on biological responses to pathogens, chemical agents, and radiactive exposures, as well as forensic goals such as attribution of samples of unknown origin. Improved proteomics technologies will advance knowledge of biological mechanisms of pathogen–chemical–radiation exposures and responses, and accelerate development of biomarkers for early detection, diagnosis, and prognosis. Of course, higher throughput, greater sensitivity, and better reproducibility are features of high salience in this domain, as in biomedicine. Contributions of proteomics to vaccine production for military populations are also highly desired. Use of proteins as biomarkers or diagnostic agents would be valuable through biometrics, biofiltering, and biodetection for surface, air, and seafaring transportation and for monitoring of foods both for national security and for general consumer safety.

5.5 Economic impact

We have learned from the development of the multiple disciplines in biotechnology and from the Human Genome Project that whole industries, new employment categories, and very large economic gains can and do result from innovation and investment in science and technology. In May 2011, the Battelle Technology Partnership Practice issued a report "Economic Impact of the Human Genome Project – How a $3.8 Billion Investment Drove $796 Billion in Economic Impact, Created 310 000 Jobs and Launched the Genomic Revolution" [1]. In 2011 dollars, that core investment is valued at $5.6B. The direct economic output during 1988–2010 was estimated at $265B of the $796B. The overall conclusion is that the payoffs can be very large for bold projects such as a Human Proteome Project. In this case, many of the hoped-for developments are still in the early stages of incubation and multiyear development pipelines. We are poised, after a decade of gestation and progress, for major impacts from proteomics.

Participation in this NSTC/NIST Workshop by scientist–managers from many federal R&D agencies showed the interest in tapping proteomics technologies and stimulating major advances in proteomics for applications in each of their sectors. Despite the current budgetary challenges, it would be desirable to invest in the scientifically promising and economically rewarding future of proteomics, including priority allocation of funds for development of next-generation technology platforms and associated informatics and databases.

There are estimates of the proteomics marketplace and projections of growth over the coming years. Front Line Strategic Consulting Inc. (San Mateo, CA, USA) projected $2.68B for the proteomics market in 2008, growing from $1.5B in 2003, according to The Industrial Physicist [29] (2005). That represented a 12% compound annual growth rate (CAGR) in the four segments of protein separation, protein characterization, content and bioinformatics, and services. Companies identified included Amersham (now GE Healthcare), Applied Biosystems Inc. (now Life Technologies), BioRad, Micromass-Waters, and Thermo-Electron. Global Industry Analysts Inc. (www.strategyr.com) projected further growth to $6.1B by 2015, according to BioMed Trends, Sept 2010, covering 168 companies, including Agilent, BioRad, Bruker Daltonics, GE Healthcare, Life Technologies, Perkin Elmer, Shimadzu, Sigma Aldrich, ThermoFisher, and Waters. These figures are indicators of high potential.

Meanwhile, there are substantial investments in competitor economies around the world. The EU has invested repeatedly in large proteomics projects, top laboratories, networks of laboratories in member countries, and data resource institutions. Sweden's government and foundations have made remarkable investments reflected in the Human Protein Atlas. China has announced very large investments; Genome Engineering News 1 September 2011

reported 250 companies in the genomics plus proteomics space in China with a CAGR of 35% moving a current market of $175M toward $655M for 2015 [30]. In 2012, the European Bioinformatics Institute announced that it will provide storage for raw mass spec data as part of its Proteomics Identifications Database, PRIDE, funded by the EU. The counterpart data repositories and data-sharing functions in the United States (Peptidome at NIH, Tranche at the University of Michigan, and PeptideAtlas at the Institute for Systems Biology) have been unable to secure sustained funding.

Proteomics is already a multibillion dollar enterprise with double-digit CAGRs. Its range of applications mirrors that of genomics and biotechnology more broadly. It is very likely to be a major contributor to job growth and economic progress, with impacts in multiple sectors.

6 Interagency opportunity to address the grand challenge in proteomics technologies

The NSTC of the White House Office of Science and Technology Policy provided the aegis for the February 2011 Gaithersburg Workshop on a Grand Challenge for Proteomics. NIST was the host, and many federal R&D agencies were represented. This report summarizes a positive view about the opportunities to exploit proteomics for the goals of each department or agency.

Interagency efforts could have four major components: (i) setting bold programmatic goals for investment in technology development, (ii) identifying research resources, (iii) pursuing biological applications, and (iv) ensuring coordination and standardization.

A multiagency effort can highlight opportunities to participate in pursuing agency-specific priorities with a coordinated R&D effort to make bold advances in proteomics technologies. A unified Grand Challenge for Proteomics can increase prospects for budget synergies for participating agencies and prospects for a large multiplier in economic activity and job growth in the relevant applied biotechnology sectors, in part through private/public partnerships. The National Bioeconomy Blueprint released in April 2012 by the White House Office of Science and Technology Policy highlighted proteomics, together with synthetic biology and bioinformatics, as "essential foundational technologies" for the bioeconomy of the future [31]. Finally, such an initiative will enhance the role of US scientists, institutions, and companies in a growing, high-profile international S&T endeavor with major societal benefits for health, agriculture, environment, energy, and national security.

Certain commercial equipment is identified in this report to adequately describe experimental procedures. Such identification does not imply recommendation or endorsement by the NIST, nor does it imply that the equipment identified is necessarily the best available for the purpose.

Note that in this document the acronym SRM always refers to Selective Reaction Monitoring and not NIST Standard Reference Materials which also use the acronym SRM.

7 References

[1] Battelle Technology Partnership Practice, *How a $3.8 Billion Investment in Human Genome Project drove $796 billion in Economic Impact Creating 310,000 Jobs and Launching the Genomic Revolution*. Battelle Memorial Institute, Washington, DC 2011.

[2] National Research Council. *A New Biology for the 21st Century*. The National Academies Press, Washington, DC 2009.

[3] National Research Council. Committee on Research at the Intersection of the Physical and Life Sciences. *Research at the Intersection of the Physical and Life Sciences*. National Academies Press., Washington, DC 2010.

[4] Lamond, A. I., Uhlen, M., Horning, S., Makarov, A. et al., Advancing cell biology through proteomics in space and time (PROSPECTS). *Mol. Cell. Proteomics* 2012, *11*, O112 017731.

[5] Farrah, T., Deutsch, E. W., Kreisberg, R., Sun, Z. et al., PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* 2012, *2*, 1170–1175.

[6] Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B. et al., Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell* 2009, *138*, 795–806.

[7] Anderson, N. L., Anderson, N. G., Haines, L. R., Hardie, D. B. et al., Mass spectrometric quantitation of peptides and proteins using stable isotope standards and capture by anti-peptide antibodies (SISCAPA). *J. Proteome Res.* 2004, *3*, 235–244.

[8] Beck, M., Schmidt, A., Malmstroem, J., Claassen, M. et al., The quantitative proteome of a human cell line. *Mol. Syst. Biol.* 2011, *7*, 549.

[9] Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N. et al., Systems-wide perturbation analysis with near complete coverage of the yeast proteome by single-shot UHPLC runs on a bench-top Orbitrap. *Mol. Cell. Proteomics* 2011, *11*, M111.013722.

[10] Chi, A., Huttenhower, C., Geer, L. Y., Coon, J. J. et al., Analysis of phosphorylation sites on proteins from *Saccharomyces cerevisiae* by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 2193–2198.

[11] Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E. et al., Towards a knowledge-based human protein atlas. *Nat. Biotechnol. 2010*, *28*, 1248–1250.

[12] Zeiler, M., Straube, W. L., Lundberg, E., Uhlen, M. et al., A protein epitope signature tag (PrEST) library allows SILAC-based absolute quantification and multiplexed

determination of protein copy numbers in cell lines. *Mol. Cell. Proteomics* 2012, *11*, O111 009613.

[13] Farrah, T., Deutsch, E. W., Omenn, G. S., Campbell, D.S. et al., A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* 2011, *10*, M110 006353.

[14] Legrain, P., Aebersold, R., Archakov, A., Bairoch, A. et al., The human proteome project: current state and future direction. *Mol. Cell. Proteomics* 2011, *10*, M111 009993.

[15] Paik, Y. K., Jeong, S. K., Omenn, G. S., Uhlen, M. et al., The chromosome-centric human proteome project for cataloguing proteins encoded in the genome. *Nat. Biotechnol.* 2012, *30*, 221–223.

[16] Vidal, M., Chan, D. W., Gerstein, M., Mann, M. et al., The human proteome – a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. *Clin. Proteomics* 2012, *9*, 6.

[17] Hood, L., A personal journey of discovery: developing technology and changing biology. *Annu. Rev. Anal. Chem.* 2008 *1*, 1–43.

[18] Omenn, G. S., AAAS presidential address: grand challenges and great opportunities in science, technology, and public policy. *Science* 2006, *314*, 1696–1704.

[19] Tate, S. A., Loboda, A., Chernushevich, I., Gillet, L. et al. (SWATH) a method for collecting MSMS of all parent ions in a sample on an LC time scale. 59th ASMS Conference. *Denver, CO.* 2011, MP 596.

[20] Agnew, H. D., Rohde, R. D., Millward, S.W., Nag, A. et al., Iterative in situ click chemistry creates antibody-like protein-capture agents. *Angew. Chem. Int. Ed. Engl.* 2009, *48*, 4944–4948.

[21] Bancaud, A., Huet, S., Rabut, G., Ellenberg, J. et al., Fluorescence perturbation techniques to study mobility and molecular dynamics of proteins in live cells: FRAP, photoactivation, photoconversion, and FLIP. *Cold Spring. Harb. Protoc.* 2010, PMID: 21123431.

[22] Rissin, D. M., Kan, C. W., Campbell, T. G., Howes, S. C. et al., Single-molecule enzyme-linked immunosorbent assay detects serum proteins at subfemtomolar concentrations. *Nat. Biotechnol.* 2010, *28*, 595–599.

[23] Prusiner, S., Cell biology. A unifying role for prions in neurodegenerative diseases. *Science* 2012, *336*, 1511–1513.

[24] Menon, R., Roy, A., Mukherjee, S., Belkin, S. et al., Functional implications of structural predictions for alternative splice proteins expressed in her2/neu-induced breast cancers. *J. Proteome Res.* 2011, *10*, 5503–5511.

[25] Institute of Medicine Committee on Omics-Based Predictive Tests, Michael, C. C., Nass, S., Omenn, G. S. (Eds.), *Evolution of Translational Omics: Lessons Learned and Path Forward*, National Academy Press, Washington DC 2012.

[26] Kinsinger, C. R., Apffel, J., Baker, M., Bian, X. et al., Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam principles). *Proteomics* 2011, DOI 10.1002/pmic.201100562; *Mol. Cell Proteomics* 2011, DOI 10.1074/mcp.0111.015446; *J. Proteome Research* 2011, DOI 10.1021/pr201071t.

[27] Hood, L., Flores, M., Personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnol.* 2012, Mar 18, doi: 10.1016/j.nbt.2012.03.004

[28] Qin, S., Zhou, Y., Lok, A. S., Tsodikov, A. et al., SRM targeted proteomics in search for biomarkers of HCV-induced progression of fibrosis to cirrhosis in HALT-C patients. *Proteomics* 2012, *12*, 1244–1252.

[29] Malsch, I., Protein research calls for advanced instruments. *Ind. Phys*, 2005, *9*, 18.

[30] Zhang, J., Omics-related research in rise in China. *Genet. Eng. Biotechnol. News* 2011, *31*, 36–37.

[31] Office of Science and Technology Policy, *National Bioeconomy Blueprint*. Office of Science and Technology Policy, Executive Office of the President, Washington DC 2012, 48 pp.

8 Addendum

Leigh Anderson, Plasma Proteome Institute; Tom Baer, Stanford Photonics Research Center; Jill Banfield, University of California, Berkeley; Maureen Beanan, NIAID, NIH; Laura Beretta, Fred Hutchinson Cancer Research Center; Jason Boehm, NIST; Mark Boggess, ARS, USDA; Judy Britz, Maryland Biotechnology Center; Richard Caprioli, Vanderbilt University; Steve Carr, Broad Institute; Kenneth Carter, Noble Life Sciences; Christine Colvis, NIDA, NIH; Robert Corn, UC Irvine; Mildred Donlon, DARPA; Charles Edmonds, NIGMS, NIH; Adam Felsenfeld, NHGRI, NIH; Catherine Fenselau, University of Maryland, College Park; Susan Fisher, UC San Francisco; Tina Gatlin, NHGRI, NIH; Yue Ge, US EPA, Research Triangle Park; Chris Geddes, University of Maryland, Baltimore County; Gradimir Georgevich, IONICS MSE, Inc; Larry Gold, Somalogic, Inc; Susan Gregurick, OBER, Dept of Energy; Ed Harlow, Harvard Medical School; Jim Heath, Caltech; Tara Hiltke, NCI, NIH; Stephen Horrigan, Hoble Life Sciences Inc; Jennie Hunter-Cevera, Research Triangle Institute; Jesseong Hwang, NIST; Joany Jackman, Applied Physics laboratory, John Hopkins University; Steven Kappes, ARS, USDA; Jim Karkanias, Microsoft Corp; Arthur Katz, OBER, Dept of Energy; laura Kiessling, University of Wisconsin-Madison; Donna Kimball, NIST; Chris Kinsinger, NCI, NIH; Mark Knepper, NHLBI, NIH; Alison Kraigsley, NIST; Josh LaBaer, Arizona State University; Brian Mansfield, Correlogic Systems Inc; Liz Mansfield, CDR/OIVD, FDA; Erica McJimpsey, NIST; Enrique Michelotti, NHGRI, NIH; Stephen Michnick, Universite de Montreal; Ken Miller, Agilent Technologies; Barbara Mittleman, OD, NIH; Rob Moritz, Institute for Systems Biology; Willie E. May, NIST; Matt Munson, NIST; Laurie Nadler, NIMH, NIH; Karen Nelson, J. Craig Venter Institute; Pankaj Oberoi, Meso Scale Discovery; Anna Palmissano, Fred Hutchinson Cancer Research Center; Amanda Paulovich, Fred Hutchinson Cancer Research Center; Scott Peterson, J. Craig Venter Institute; Karen Phinney, NIST; Rembert Pieper, J. Craig Venter Institute; Anne

Plant, NIST; Melanie Roberts, University of Colorado, Boulder; Henry Rodriguez, NCI, NIH; Chuck Romine, NIST; Paul Rudnick, NIST; Marc Salit, NIST; Peter Schad,Research Triangle Institute; John Schiel, NIST; Salvatore Sechi, NIDDK, NIH; Doug Sheeley, NCRR, NIH; Richard Smith, Pacific Northwest National Laboratory; Pothur Srinivas, NHLBI, NIH; Michael Stebbins, Office of Science and Technology Policy; Stephen Stein, NIST; Zivana Tezak, OIVD/CDRH, FDA; David Thompson, Dept of Energy; Roger Tsien, UC San Diego; Lili Wang, NIST; Witold Winick, US EPA; June Wispelwey, AIChE; Cathy Wu, University of Delaware; Rebecca Zangmeister, NIST.

